Studies in Classification, Data Analysis, and Knowledge Organization

Titles in the Series

E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy (Eds.)
New Approaches in Classification and Data Analysis. 1994 (out of print)

W. Gaul and D. Pfeifer (Eds.)
From Data to Knowledge. 1995

H.-H. Bock and W. Polasek (Eds.)
Data Analysis and Information Systems. 1996

E. Diday, Y. Lechevallier, and O. Opitz (Eds.)
Ordinal and Symbolic Data Analysis. 1996

R. Klar and O. Opitz (Eds.)
Classification and Knowledge Organization. 1997

C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, and Y. Baba (Eds.)
Data Science, Classification, and Related Methods. 1998

I. Balderjahn, R. Mathar, and M. Schader (Eds.)
Classification, Data Analysis, and Data Highways. 1998

A. Rizzi, M. Vichi, and H.-H. Bock (Eds.)
Advances in Data Science and Classification. 1998

M. Vichi and O. Opitz (Eds.)
Classification and Data Analysis. 1999

W. Gaul and H. Locarek-Junge (Eds.)
Classification in the Information Age. 1999

H.-H. Bock and E. Diday (Eds.)
Analysis of Symbolic Data. 2000

H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, and M. Schader (Eds.)
Data Analysis, Classification, and Related Methods. 2000

W. Gaul, O. Opitz, and M. Schader (Eds.)
Data Analysis. 2000

R. Decker and W. Gaul (Eds.)
Classification and Information Processing at the Turn of the Millenium. 2000

S. Borra, R. Rocci, M. Vichi, and M. Schader (Eds.)
Advances in Classification and Data Analysis. 2001

W. Gaul and G. Ritter (Eds.)
Classification, Automation, and New Media. 2002

K. Jajuga, A. Sokołowski, and H.-H. Bock (Eds.)
Classification, Clustering and Data Analysis. 2002

M. Schwaiger and O. Opitz (Eds.)
Exploratory Data Analysis in Empirical Research. 2003

M. Schader, W. Gaul, and M. Vichi (Eds.)
Between Data Science and Applied Data Analysis. 2003

H.-H. Bock, M. Chiodi, and A. Mineo (Eds.)
Advances in Multivariate Data Analysis. 2004

D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul (Eds.)
Classification, Clustering, and Data Mining Applications. 2004

D. Baier and K.-D. Wernecke (Eds.)
Innovations in Classification, Data Science, and Information Systems. 2005

M. Vichi, P. Monari, S. Mignani, and A. Montanari (Eds.)
New Developments in Classification and Data Analysis. 2005

D. Baier, R. Decker, and L. Schmidt-Thieme (Eds.)
Data Analysis and Decision Support. 2005

C. Weihs and W. Gaul (Eds.)
Classification – the Ubiquitous Challenge. 2005

M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul (Eds.)
From Data and Information Analysis to Knowledge Engineering. 2006

V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna (Eds.)
Data Science and Classification. 2006

Sergio Zani · Andrea Cerioli
Marco Riani · Maurizio Vichi
Editors

Data Analysis, Classification and the Forward Search

Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6–8, 2005

With 118 Figures and 50 Tables

Springer

Prof. Sergio Zani
Department of Economics
Section of Statistics
and Computing
University of Parma
Via Kennedy 6
43100 Parma, Italy
sergio.zani@unipr.it

Prof. Marco Riani
Department of Economics
Section of Statistics
and Computing
University of Parma
Via Kennedy 6
43100 Parma, Italy
mriani@unipr.it

Prof. Andrea Cerioli
Department of Economics
Section of Statistics
and Computing
University of Parma
Via Kennedy 6
43100 Parma, Italy
andrea.cerioli@unipr.it

Prof. Maurizio Vichi
Department of Statistics,
Probability and Applied
Statistics
University of Rome
"La Sapienza"
Piazzale Aldo Moro 5
00185 Roma, Italy
maurizio.vichi@uniroma1.it

ISSN 1431-8814
ISBN 10 3-540-35977-X Springer Berlin Heidelberg New York
ISBN 13 978-3-540-35977-7 Springer Berlin Heidelberg New York

Cover: Erich Kirchner, Heidelberg
Production: LE-TEX, Jelonek, Schmidt & Vöckler GbR, Leipzig

SPIN 11789703 Printed on acid-free paper – 43/3100 – 5 4 3 2 1 0

Preface

This volume contains revised versions of selected papers presented at the biennial meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, which was held in Parma, June 6-8, 2005. Sergio Zani chaired the Scientific Programme Committee and Andrea Cerioli chaired the Local Organizing Committee.

The scientific programme of the conference included 127 papers, 42 in specialized sessions, 68 in contributed paper sessions and 17 in poster sessions. Moreover, it was possible to recruit five notable and internationally renowned invited speakers (including the 2004–2005 President of the International Federation of Classification Societies) for plenary talks on their current research work. Among the specialized sessions, two were organized by Wolfgang Gaul with five talks by members of the GfKl (German Classification Society), and one by Jacqueline J. Meulman (Dutch/Flemish Classification Society). Thus, the conference provided a large number of scientists and experts from home and abroad with an attractive forum for discussion and mutual exchange of knowledge. The topics of all plenary and specialized sessions were chosen to fit, in the broadest possible sense, the mission of CLADAG, the aim of which is "to further methodological, computational and applied research within the fields of Classification, Data Analysis and Multivariate Statistics".

A peer-review refereeing process led to the selection of 46 extended papers, which are contained in this book. The more methodologically oriented papers focus on developments in clustering and discrimination, multidimensional data analysis, data mining, and robust statistics with a special emphasis on the novel Forward Search approach. Many papers also provide significant contributions in a wide range of fields of application. Customer satisfaction and service evaluation are two examples of such emerging fields. This suggested the presentation of the 46 selected papers in six parts as follows:

1. Clustering and Discrimination
2. Multidimensional Data Analysis and Multivariate Statistics
3. Robust Methods and the Forward Search
4. Data Mining Methods and Software
5. Multivariate Methods for Customer Satisfaction and Service Evaluation
6. Multivariate Methods in Applied Science

We wish to express our gratitude to the other members of the Scientific Programme Committee

B. Chiandotto, N.C. Lauro, P. Monari, A. Montanari, C. Provasi, G. Vittadini

and to the specialized session organizers

F. Camillo, M. Chiodi, W. Gaul, S. Ingrassia, J.J. Meulman

for their ability to attract interesting contributions, and to the authors, whose enthusiastic participation made the meeting possible. We would also like to extend our thanks to the chairpersons and discussants of the sessions for their stimulating comments and suggestions. We are very grateful to the referees for their careful reviews of all submitted papers and for the time spent in this professional activity.

We gratefully acknowledge the University of Parma and its Department of Economics for financial support and hospitality. We are also indebted to *Istat - Istituto Nazionale di Statistica* and *SAS* for their support.

We thank all the members of the Local Organizing Committee

A. Corbellini, G. Gozzi, L. Grossi, F. Laurini, M.A. Milioli, G. Morelli, I. Morlini

for their excellent work in managing the organization of the CLADAG-2005 conference. Special thanks go to Prof. Isabella Morlini, for her skilful accomplishment of the duties of Scientific Secretary of CLADAG-2005, and to Dr. Fabrizio Laurini for his assistance in producing this volume.

Finally, we would like to thank Dr. Martina Bihn of Springer-Verlag, Heidelberg, for her support and dedication to the production of this volume.

Parma and Rome,
June 2006

Sergio Zani
Andrea Cerioli
Marco Riani
Maurizio Vichi

Contents

Part I

Clustering and Discrimination

Genetic Algorithms-based Approaches for Clustering Time Series

Roberto Baragona[1] and Salvatore Vitrano[2]

[1] Department of Sociology and Communication,
University of Rome "La Sapienza", Italy
roberto.baragona@uniroma1.it
[2] Statistical Office,
Ministry for Cultural Heritage, Italy
svitrano@beniculturali.it

Abstract. Cluster analysis is to be included among the favorite data mining techniques. Cluster analysis of time series has received great attention only recently mainly because of the several difficult issues involved. Among several available methods, genetic algorithms proved to be able to handle efficiently this topic. Several partitions are considered and iteratively selected according to some adequacy criterion. In this artificial "struggle for survival" partitions are allowed to interact and mutate to improve and produce a "high quality" solution. Given a set of time series two genetic algorithms are considered for clustering (the number of clusters is assumed unknown). Both algorithms require a model to be fitted to each time series to obtain model parameters and residuals. These methods are applied to a real data set concerned with the visitors flow recorded, in state owned museums with paid admission, in the Lazio region of Italy.

1 Introduction

Clustering time series, that is division of time series into homogeneous subgroups, is composed of several steps. First, pre-processing is almost always needed for removing or softening unwanted characteristics that may bias the analysis (for instance, outliers, missing observations and Easter and trading day effects). Moreover, adjustment for seasonality and trend could possibly be required to allow some methods to run properly. Among many available methods X-12 ARIMA (Findley et al. (1998)) and Tramo-Seats (Gómez and Maravall (1996)) are currently adopted by Statistical Authorities in many Countries. They are well founded on theoretical grounds and supported by computer programs that make easier their application.

Then, the extraction of measurements may take place so that either the usual matrix units (time series) per variables (the measurements) or a matrix of distances between each pair of time series is available. Liao (2005) in a comprehensive survey distinguishes whether methods are based on the raw data directly, on features extracted from the data or on models built on the data.

Third step, choosing the cluster method, is closely related to the preceding step as the method largely depends on the available data structure. Four main classes may be distinguished, that is partitioning methods, hierarchical methods, density-based clustering and grid-based clustering (see, for instance, Berkhin (2002) for a comprehensive survey).

The fourth step is concerned with the choice of the algorithm. As clustering problems arise in so many fields (these range from sociology and psychology to commerce, biology and computer science) the implementation and design of algorithms continue to be the subject of active research. Over the last two decades a new class of algorithms has been developed, namely the optimization heuristics. Examples are evolutionary algorithms (simulated annealing, threshold accepting), neural networks, genetic algorithms, tabu search, ant colony optimization (see, for instance, Winker and Gilli (2004)). Optimization heuristics may cope with problems of high complexity whose potential solutions are a large discrete set. This is the case of the admissible partitions of a set of time series. In addition, assumptions on the form of the final partition (either hard or fuzzy assignments), for instance, or on the number of clusters (either known or unknown) are easily handled by optimization heuristics and require only slight modifications of the basic algorithm. Clustering time series by meta heuristic methods was investigated by Baragona (2001) while Pattarin et al. (2004) examined genetic algorithms-based approaches.

In this paper genetic algorithms (GAs) are used for implementing two model-based-methods, the first one based on the cross correlations (Zani (1983)), the second one based on the autoregressive distance (Piccolo (1990)). Other optimization heuristics may be of use, but GAs seem to ensure most flexibility and vast choice to meet the special requirements involved in clustering time series. Both algorithms are tested on the data set of the visitors of museums, monuments and archaeological sites in the Lazio region of Italy.

The rest of the paper is organized as follows. The next Section includes a description of the two clustering time series methods and the GAs are described. Results of the application to the real data set are displayed in Section 3. Section 4 concludes.

2 Clustering Methods and Genetic Algorithms

Given a set of time series two methods are considered for clustering. Both methods require a model to be fitted to each time series to obtain model parameters and residuals. The number of clusters g is assumed unknown. The fitted models are autoregressive integrated moving-average (ARIMA) models (Box et al. (1994)). The first method is aimed at grouping together time series according to the residuals cross correlations. The second method is aimed at grouping together time series that share a similar model's structure.

Let the time series $\{x_t\}$ be generated by the ARIMA$(p,d,q)(P,D,Q)_s$ model

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^d x_t = \theta(B)\Theta(B^s)a_t, \tag{1}$$

where $\{a_t\}$ is a white noise process with finite variance σ^2. The polynomials in (1) are defined

$$\phi(B)\Phi(B^s) = (1-\phi_1 B - \ldots - \phi_p B^p)(1-\Phi_1 B^s - \ldots - \Phi_P B^{Ps}) \tag{2}$$

$$\theta(B)\Theta(B^s) = (1-\theta_1 B - \ldots - \theta_q B^q)(1-\Theta_1 B^s - \ldots - \Theta_Q B^{Qs}), \tag{3}$$

and have not common factors. B is the back-shift operator, that is $B^k x_t = x_{t-k}$. Model (1) is stationary if the polynomial (2) has all roots outside the unit circle and is invertible if the polynomial (3) has all roots outside the unit circle. Under this latter assumption, the time series $\{x_t\}$ admits the infinite autoregressive representation $x_t = \Sigma\pi_k x_{t-k}$. The coefficients π_k are called π-weights.

For the first method a cluster is required to satisfy the following condition (Zani (1983)). Given a set of k time series $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n})$, $i = 1, \ldots, k$, a subset C which includes k' series ($k' < k$) is said to form a group if, for each of the $k'(k'-1)/2$ residuals cross-correlations $\rho_{i,j}(\tau)$, we have

$$|\rho_{i,j}(\tau)| > c(\alpha) \tag{4}$$

for at least a lag τ between $-m$ and m, and $i, j \in C, i \neq j$. A positive integer m has to be pre-specified which denotes the maximum lag. If all time series have n as a common number of observations, then choosing the significance level $\alpha = 0.05$, say, gives the figure $c(\alpha) = 1.96/\sqrt{n}$ in (4). The previously stated definition does not exclude that a time series may belong to more than a single group. Then there are possibly several allowable partitions to consider, and their number may be very large. Meta heuristic methods, in particular GAs, were proposed by Baragona (2001) to find the best feasible partition. As an overall objective function a modification of the k-min cluster criterion (Sahni and Gonzalez (1976)) was assumed

$$f^+(C_1, C_2, \ldots, C_g; g) = \sum_{\omega=1}^{g} \sum_{i,j \in C_\omega, i \neq j} d_{i,j}^+, \tag{5}$$

where (5) has to be maximized and

$$d_{i,j}^+ = \max\{|\rho_{i,j}(\tau)|\}, \qquad \tau = -m, \ldots, m. \tag{6}$$

When using (5), it is crucial that each cluster be a group, according to (4), for, otherwise, any algorithm, unless prematurely ended, will put together all time series into a single cluster.

The second method is new and has been developed along the same guidelines, except that the autoregressive distance proposed by Piccolo (1990) is

adopted instead of the cross-correlations-based dissimilarity index. Each time series $\{x_t, t = 1, 2, \ldots n\}$ is associated to the first m π-weights of the autoregressive representation of the ARIMA model. The first m π-weights may be computed from the coefficients of the ARIMA model (1) using the equations given in Box et al. (1994). The positive integer m is a truncation point that has to be pre-specified. For each time series there is a set of measurements $\pi_{v,1}, \pi_{v,2}, \ldots \pi_{v,m}, v = 1, 2, \ldots k$ that allows clusters to be determined. The π-weights define the autoregressive distance

$$d_{i,j} = \sqrt{\sum_{h=1}^{m} (\pi_{i,h} - \pi_{j,h})^2}. \tag{7}$$

The distributional properties of the autoregressive distance (7) were studied by Piccolo (1990) under the assumption that the time series are uncorrelated. The presence of correlation between time series was shown (Corduas (1992)) to modify the distribution of (7). In Corduas (2000) an approximation is provided that allows a formal test to be established. The squared autoregressive distance $d_{i,j}^2$ is approximately distributed as a random variable $a\chi_\nu^2 + b$ where χ_ν^2 is the chi-squared random variable with ν degrees of freedom. The constants a, b and ν depend on the common (under the null hypothesis that the time series $\{x_{i,t}\}$ and $\{x_{j,t}\}$ are generated by the same ARIMA model) variance-covariance matrix of the estimated π-weights. The condition for a set of time series to form a group has to be re-formulated in terms of the approximate critical values that may be computed for the squared distance (7). Two time series are allowed to be included in the same cluster if their squared distance is less than $a\chi_\nu^2(\alpha) + b$, where $\chi_\nu^2(\alpha)$ is the $(1 - \alpha)$-quantile of the chi-squared distribution with ν degrees of freedom and α is the significance level.

It may be argued, extending the results reported by Corduas (1992), that if the cross correlations $\rho_{i,j}$ are close to unity the two methods are likely to yield similar results. Note that Tong and Dabas (1990) include the index (6), with $\tau = 0$, among the measures of dissimilarity for a set of time series models though no threshold-based constraints were introduced.

GAs were introduced by Holland (1975) to provide evolutionary models for the adaptation process to the environment of individuals belonging to a given population. If the evolution of the best fit individual is recorded, GAs may be viewed as optimization tools. In this case, a numerical measure is used to evaluate the adaptation to the environment. This measure is called fitness function and it is the objective function which has to be maximized. The fitness function is not required to possess special mathematical properties but to be positive and non-decreasing function of the adaptation to the environment. Each potential solution to the optimization problem has to be coded as a string of ℓ characters, for instance a binary string of length ℓ (this string is usually called chromosome). There is no need to enumerate all

solutions, which has to be considered fairly impossible, but only a set of rules by which any string may be decoded in a meaningful way and assigned a positive real number. In practice GAs are useful because they allow very large spaces to be searched for solutions and very mild assumptions are required for the objective function. For a detailed description of GAs see, for instance, Goldberg (1989) and Haupt and Haupt (2004). Convergence properties have been discussed by Reeves and Rowe (2003), among others.

GAs start with an initial population of candidate solutions (individuals) that are said to form a population though they are actually a sample from the set of potential solutions. The population evolves through a iterative procedure. Each iteration usually includes three steps, that is selection, crossover and mutation. Selection is aimed at choosing the individuals with high fitness. Selection is done with replacement, so that many copies of the same individual may enter the next generation. The chromosomes of the selected individuals are possibly combined, by means of the crossover operator, to produce new individuals. Mutation of some characters within the genetic pool may take place with usually small probability, 0.001 for instance. The generational gap is the fraction of the old population that is replaced by new individuals. If the new population entirely replaces the past one the generational gap is equal to unity. Iterations continue either after a pre-specified number of generations or when a stopping criterion is met. Often the elitist strategy is used, that is the best fit individual found in a iteration is inserted in the next generation unless a better individual is found (see, for instance, Jennison and Sheehan (1995)).

Both methods are implemented by GAs with permutation encoding, often called ordered GAs (Jones and Beltramo (1991)). Time series are assigned labels $1, 2, \ldots k$, and several random permutations are considered. For each one, a random number of cluster g is generated and the g time series at the top of the list are assumed as cluster centers. Then, aggregation of the remaining time series to the nearest center is performed. To improve the solutions, the permutations are evolved through some iterations by the GAs operators selection, crossover and mutation. These have to be specially designed according to chosen encoding method and distance measure. The ordered GAs proved to be very effective in practice as far as clustering procedures are concerned.

3 Application to Real Data

The number of visitors to cultural sites and state owned museums (according to the definition given by the International Council of Museums (ICOM)), with admission fees, in the Lazio region of Italy is collected monthly by the Statistical Office at the Ministry for Cultural Heritage. We considered 37 time series from January 1996 to December 2003. Adjustment for outliers and missing data, and ARIMA model (1) building were performed by the computer program Tramo-Seats (Gómez and Maravall (1996)). This program may be

downloaded from the web site http://www.istat.it. All series were reduced to have common number of observations 78. Two time series were discarded because the sites happened to be closed many times during the observation period. For all remaining 35 time series the ARIMA coefficients were used to compute the π-weights. Also, residuals from ARIMA models were recorded and the cross correlations were computed. All estimated ARIMA models used for cluster analysis passed the Ljung-Box test (based on the first 24 residuals autocorrelations) at the 5% significance level (at the 1% significance level for series 10).

The first method groups series with cross correlations absolute values greater than 0.2219 (1.96 divided by the square root of the number of observations). The clusters that have been formed are reported in Table 1 (series are numbered from 1 to 35).

Cluster	**Time series**
(1)	2 5 10 11 14 15 21 22
(2)	19 23 24 25 26 29 32 35
(3)	6 7 27 30 31
(4)	8 9 16 17 18 20
(5)	4 12 13 28 33
(6)	1 34
(7)	3

Table 1. Clusters of cultural sites visitors in Lazio, Italy (first method)

The second method groups the time series according to the π-weights. Time series may be grouped together only if their pairwise squared autoregressive distance is less than a threshold whose values for this data set have been found to vary in the interval $[0.12, 0.17]$ (for this method thresholds for acceptance vary through time series pairs). The two methods are likely to yield similar results if the cross correlations are large. This is not the present case, however, as most cross correlations are less than 0.3, some are between 0.3 and 0.5 and only one cross correlation exceeds 0.5 (it is about 0.6). We obtained 9 clusters that are reported in Table 2. First and second clusters contain 23 out of 35 series. Clusters 3 and 6 are very small. The first one includes series that are similar to series in clusters 1 and 2, while the second one contains series that are considerably different. As far as time series 26, 34, 12, 17 and 23 are concerned, each one forms a single cluster.

The first method seems to cluster together time series according to spatial and typological closeness. Cluster 1 includes museums located in Rome. Cluster 2 include archaeological areas in Rome too. Cluster 3 includes archaeological areas as well, but out of Rome. The sites that belong to the historic / artistic local authority are in cluster 4. Evidence of typological clustering is provided by cluster 5, which includes four sites mostly archae-

Cluster	Time series
(1)	3 4 6 7 11 13 15 28 30 32 35
(2)	1 2 9 10 16 18 19 20 21 24 25 33
(3)	5 27 31
(4), (5)	26, 34
(6)	8 14 22 29
(7), (8), (9)	12, 17, 23

Table 2. Clusters of cultural sites visitors in Lazio, Italy (second method)

ological museums. The second method seems to group time series according to sites easy to access as perceived by visitors. This is a new interesting issue that emerges from cluster analysis. Foreign tourists, for instance, are unlikely to visit remote sites while it is easy for a school to visit sites nearby. Also, scholars may consider important to visit sites though difficult to access. These circumstances produce different time series dynamic behaviors. The first cluster includes most archaeological sites out of Rome that are likely to attract mostly local public. Cluster 3 is very similar. Cluster 2 includes many museums in Rome that foreign tourists, for instance, hardly miss to visit. Clusters with single time series are explained by some special characteristics of the site. Cluster 4, for example, includes only Villa d'Este - Tivoli. Unlike most sites out of Rome, this site attracts a rather regular visitors flow. Another example is cluster 5 that includes the archaeological tour: Colosseum, Palatine Hill and Forum. This site too is peculiar as it sells combined tickets, that is tickets that allow the tourist to visit these and other sites.

4 Concluding Remarks

Genetic algorithms were designed to implement two methods for clustering time series. The first one is based on the residuals cross correlations, the second one on the time series models structures. As an example, the two methods are applied for clustering the time series of the visitors to the museums, monuments and archaeological areas of the Lazio region of Italy. The first method seems to group time series according to the spatial locations of the sites. The second method seems to group the time series together according to both visitors typology and sites characteristics. Cross correlations are rather small (though larger than the critical values) and the two methods are expected to produce different results in this case. Performance of genetic algorithms may be considered quite good. Computations were fast and the results seem to be reliable and accurate.

References

BARAGONA, R. (2001): A Simulation Study on Clustering Time Series with Meta-Heuristic Methods. *Quaderni di Statistica, 3, 1–26.*

BERKHIN, P. (2002): *Survey of Clustering Data Mining Techniques.* Technical Report, Accrue Software, San José, California, http://citeseer.ist.psu.edu/berkhin02survey.html.

BOX, G.E.P., JENKINS, G.M. and REINSEL, G.C. (1994): *Time Series Analysis. Forecasting and Control (3rd Edition).* Prentice Hall, San Francisco.

CORDUAS, M. (1992): Una Nota sulla Distanza tra Modelli ARIMA per Serie Storiche Correlate. *Statistica, 52, 515–520.*

CORDUAS, M. (2000): La Metrica Autoregressiva tra Modelli ARIMA: una Procedura in Linguaggio GAUSS. *Quaderni di Statistica, 2, 1–37.*

FINDLEY, D.F., MONSELL,B.C., BELL, W.R., OTTO, M.C. and CHEN, B.-C. (1998): New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program (with discussion). *Journal of Business and Economic Statistics, 16, 127–176.*

GOLDBERG, D.E. (1989): *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, Massachusetts.

GÓMEZ, V. and MARAVALL, A. (1996): *Programs Tramo and Seats: Instructions for Users.* Technical Report 9628, The Banco de España, Servicios de Estudios.

HAUPT, R.L. and HAUPT, S.E. (2004): *Practical Genetic Algorithms (2nd Edition).* John Wiley & Sons, Hoboken, New Jersey.

HOLLAND, J.H. (1975): *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor.

JENNISON, C. and SHEEHAN, N. (1995): Theoretical and empirical properties of the genetic algorithm as a numerical optimizer, *Journal of Computational and Graphical Statistics, 4, 296-318.*

JONES, D.R. and BELTRAMO, M.A. (1991): Solving Partitioning Problems with Genetic Algorithms. In: R.K. Belew and L.B. Booker (Eds.): *Proceedings of the Fourth International Conference on Genetic Algorithms.* Morgan Kaufmann, San Diego, California, 442–449.

LIAO, T.W. (2005): Clustering of Time Series Data–A Survey. *Pattern Recognition, 38, 1857–1874.*

PATTARIN, F., PATERLINI, S. and MINERVA, T. (2004): Clustering Financial Time Series: an Application to Mutual Funds Style Analysis. *Computational Statistics & Data Analysis, 47, 353–372.*

PICCOLO, D. (1990): A Distance Measure for Classifying ARIMA Models. *Journal of Time Series Analysis, 11, 153–164.*

REEVES, C.R. and ROWE, J.E. (2003): *Genetic Algorithms - Principles and Perspective: a Guide to GA Theory.* Kluwer Academic Publishers, London.

SAHNI, S. and GONZALEZ, T. (1976): P-Complete Approximation Problems. *Journal of the Association for Computing Machinery, 23, 555–565.*

TONG, H. and DABAS, P. (1990): Cluster of Time Series Models: an Example. *Journal of Applied Statistics, 17, 187–198.*

WINKER, P. and GILLI, M. (2004): Application of Optimization Heuristics to Estimation and Modelling Problems. *Computational Statistics & Data Analysis, 47, 211–223.*

ZANI, S. (1983): Osservazioni sulle Serie Storiche Multiple e l'Analisi dei Gruppi. In: D. Piccolo (Ed.): *Analisi Moderna delle Serie Storiche.* Franco Angeli, Milano, 263–274.

On the Choice of the Kernel Function in Kernel Discriminant Analysis Using Information Complexity

Hamparsum Bozdogan[1], Furio Camillo[2], and Caterina Liberati[2]

[1] Department of Statistics, Operations, and Management Science
The University of Tennessee
Knoxville, TN 37996-0532 U.S.A.
bozdogan@utk.edu
[2] Dipartimento di Scienze Statistiche
Università di Bologna, Italy
{fcamillo, liberati}@stat.unibo.it

Abstract. In this short paper we shall consider the Kernel Fisher Discriminant Analysis (KFDA) and extend the idea of Linear Discriminant Analysis (LDA) to nonlinear feature space. We shall present a new method of choosing the optimal kernel function and its effect on the KDA classifier using information-theoretic complexity measure.

1 Introduction and the Problem

Discriminant analysis (DA) is one of the popular multivariate methods which has a long history. DA is a classification problem that consists of assigning or classifying an individual or object to one of several known or unknown alternative classes (or groups) on the basis of many measurements on the individuals, objects, or cases. The goal of discriminant analysis is: given a data set with two or more than two classes (or groups), say, find the best feature or feature set either linear or non-linear to discriminate between the classes and maximize average class separation. Equivalently, we attempt to minimize the probability of missclassification.

Recently in statistical data mining and knowledge discovery, kernel-based methods have attracted attention from many researchers. As a result, many kernel-based methods have been developed. They have become popular tools for classification, clustering, and regression analysis in the machine learning community since the introduction of support vector machines (SVMs) during the early 1990s. The popularity of the method stems from the fact that kernel methods almost always outperform traditional multivariate statistical techniques. Now, we can carry out kernel based approaches to all the classical multivariate procedures. Examples of these include, kernel principal component analysis (KPCA), kernel logistic regression (KLR), kernel Fisher discriminant analysis (KFDA), or in short kernel discriminant analysis (KDA), kernel canonical correlations (KCC), etc., to mention a few. These

methods are characterized by transformation of the input data to a high dimensional feature space, followed by application of the technique in question to the transformed data.

In this paper we shall consider Kernel Fisher Discriminant Analysis (KFDA) and extend the idea of Linear Discriminant Analysis (LDA) to nonlinear feature space. We shall present a new method of choosing the optimal kernel function and explore its effect on the KDA classifier. In general the problem of which is the most appropriate kernel for a particular real application or problem is still an open problem in the literature. In this short paper, we will introduce a new special form of the information-theoretic measure of complexity of Bozdogan (1988, 1990, 1994, 2000, 2004) to choose the optimal kernel function.

We will illustrate our result using a toy example on a benchmark data set of Ripley (1994) and discuss future work on model selection in kernel methods.

2 Kernel Discriminant Analysis (KDA)

Reproducing Kernel Hilbert Space (RKHS) were developed by Aronszajn in 1950. A *RKHS* is defined by a positive definite kernel function

$$K : \mathcal{R}^d \times \mathcal{R}^d \rightarrow \mathcal{R} \tag{1}$$

on pairs of points in data space.

If these kernel functions satisfy the Mercer's condition (Mercer, 1909, Cristianini and Shawe-Taylor, 2000), they correspond to non-linearly mapping the data to a higher dimensional *feature* space $\mathcal{F}$ by a map

$$\Phi : \mathcal{R}^d \rightarrow \mathcal{F} \tag{2}$$

and taking the dot product in this space (Vapnik, 1995):

$$K(x, y) = \Phi(x) \cdot \Phi(y). \tag{3}$$

This means that any linear algorithm in which the data only appears in the form of dot products $< \mathbf{x}_i, \mathbf{x}_j >$ can be made nonlinear by replacing the dot product by the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ and doing all the other calculations as before. In other words, each data point is mapped nonlinearly to a higher dimensional feature space.

As is well-known the Fisher Linear Discriminant analysis (FLDA) or in short LDA, is one of the most frequently used classification techniques. In order to make LDA applicable to nonlinear data in a feature space induced by a Mercer kernel, we need to develop and utilize kernel methods also referred to *"kernel machines"*. This approach gives rise to a nonlinear pattern recognition method which has very impressive performance on real data sets.

Assume that we are given the input data set $\mathcal{I}_{XY} = \{(x_1, y_1), ..., (x_n, y_n)\}$ of training vectors $\mathbf{x}_i \in \chi \subseteq \mathcal{R}^d$ and the corresponding values of $y_i \in \mathcal{Y} = \{1, 2\}$. The y_i are sets of indices of training vectors belonging to the first $y = 1$ and the second $y = 2$ class, respectively. The class separability in a direction of the weights $\alpha = [\alpha_1, \ldots, \alpha_n]'$ in the *feature space* $\mathcal{F}$ is defined such that the Fisher criteria:

$$J_F(w) = \frac{\alpha' S_B^{\Phi} \alpha}{\alpha' S_W^{\Phi} \alpha}, \tag{4}$$

is maximized, where S_B^{Φ}, S_W^{Φ} are respectively the *between and within covariance matrices* in the future space. That is,

$$\begin{aligned} S_B^{\Phi} &= (m_1^{\Phi} - m_2^{\Phi})(m_1^{\Phi} - m_2^{\Phi})^T \\ &= (\overline{\kappa}_1 - \overline{\kappa}_2)(\overline{\kappa}_1 - \overline{\kappa}_2')', \end{aligned} \tag{5}$$

$$\begin{aligned} S_W^{\Phi} &= \sum_{i=1,2} \sum_{x \in X_i} (\Phi(x) - m_i^{\Phi})(\Phi(x) - m_i^{\Phi})^T \\ &= KK' - \sum_{k=1}^{2} n_k \overline{\kappa}_k \overline{\kappa}_k' \end{aligned} \tag{6}$$

with

$$K = [\kappa(x_i, x_j)]_{(n \times n)}, \text{and}$$

$$\overline{\kappa}_k = \frac{1}{n_k} \sum_{j \in I_k} K_j,$$

where K_j is the j-th column of K and I_k the index set of group k.

The kernel discriminant function $f(x)$ of the binary classifier

$$q(\mathbf{x}) = \begin{cases} 1 & for\ f(x) \geq 0, \\ 2 & for\ f(x) < 0 \end{cases} \tag{7}$$

can be written as

$$\begin{aligned} f_y(x) &= \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) + b_y \\ &= < \alpha_y, \kappa(\mathbf{x}) > + b_y, \quad y \in \mathcal{Y}. \end{aligned} \tag{8}$$

With α being solved from (4), the intercept (or the bias) b of the discriminant hyperplane (8) is determined by forcing the hyperplane to pass through the mid point of the two group means. That is,

$$b = -\alpha' \frac{(\overline{\kappa}_1 + \overline{\kappa}_2)}{2}. \tag{9}$$

3 Regularized Kernel Discriminant Analysis (RKDA)

Without loss of generality, dropping the superscript of S_W^{Φ} and S_B^{Φ}, the coefficients α are given by the leading eigenvector of $S_W^{-1}S_B$.

Since the matrix S_W is at most of rank $n-1$, it is not strictly positive and can even fail to be positive semi-definite due to numerical problems.

Therefore, we regularize it by adding a penalty function μI to overcome the numerical problem caused by singular within-group covariance S_W. In this case, the criterion

$$J_F(w) = \frac{\alpha' S_B \alpha}{\alpha'(S_W + \mu I)\alpha} \tag{10}$$

is maximized, where the diagonal matrix μI in the denominator of the criterion (10) serves as the regularization term. If μ is sufficiently large, then S_W is numerically more stable and becomes positive definite. Another possible regularization would be to add a multiple of the kernel matrix K to S_W as suggested by Mika (2002, p. 46). That is,

$$S_W(\mu) = S_W + \mu K, \quad \mu \geq 0, \tag{11}$$

but this does not work well in practice.

If we let the covariance matrix $\hat{\Sigma}_W$

$$\hat{\Sigma}_W = \frac{1}{n} S_W, \tag{12}$$

then $\hat{\Sigma}_W$ degenerates when the data dimension p increases. In cases when the number of variables p is much larger than the number of observations n, and in general, it makes sense to utilize improved methods of estimating the covariance matrix Σ_W. We call these estimators *smoothed*, *robust*, or *stoyki* covariance estimators. These are given as follows.

- The *stipulated diagonal covariance estimator (SDE)*:

$$\hat{\Sigma}_{SDE} = (1-\pi)\hat{\Sigma}_W + \pi Diag(\hat{\Sigma}_W), \tag{13}$$

where $\pi = p(p-1)\left[2n\left(trR^{-1} - p\right)\right]^{-1}$ and where

$$R = (Diag(\hat{\Sigma}_W))^{-1/2}\hat{\Sigma}_W(Diag(\hat{\Sigma}_W))^{-1/2} \tag{14}$$

is the correlation matrix.

The SDE estimator is due to Shurygin (1983). SDE avoids scale dependence of the units of measurement of the variables.

- The *convex sum covariance estimator (CSE)*:

Based on the quadratic loss function used by Press (1975), Chen (1976) proposed a *convex sum covariance matrix estimator (CSE)* given by

$$\hat{\Sigma}_{CSE} = \frac{n}{n+m}\hat{\Sigma}_W + (1 - \frac{n}{n+m})\hat{D}_W, \tag{15}$$

where $\hat{D}_W = (\frac{1}{p}tr\hat{\Sigma}_W)\mathbf{I}_p$. For $p \geq 2$, m is chosen to be

$$0 < m < \frac{2[p(1+\beta) - 2]}{p - \beta}, \tag{16}$$

where

$$\beta = \frac{(tr\hat{\Sigma}_W)^2}{tr(\hat{\Sigma}_W^2)}. \tag{17}$$

This estimator improves upon $\hat{\Sigma}_W$ by shrinking all the estimated eigenvalues of $\hat{\Sigma}_W$ toward their common mean. Note that there are other smoothed covariance estimators. For space considerations, we will not discuss them in this paper. For more on these, see Bozdogan (2006).

4 Choice of Kernel Functions

One of the important advantages of kernel methods, including KDA, is that the optimal model parameters are given by the solution of a convex optimization problem with a single, global optimum. However, optimal generalization still depends on the selection of a suitable kernel function and the values of regularization and kernel parameters. See, e.g., Cawley and Talbot (2003, p. 2).

There are many kernel functions to choose from. The most common kernel functions are *Gaussian RBF* ($c \in \mathcal{R}$), *polynomial* ($d \in \mathcal{N}, c \in \mathcal{R}$), *sigmodial* ($a, b \in \mathcal{R}$), *PE kernel* ($r \in \mathcal{R}, \beta \in \mathcal{R}_+$), *Cauchy kernel* ($c \in \mathcal{R}_+$), and *inverse multi-quadric* ($c \in \mathcal{R}_+$) kernel functions are among the most common ones.

The main idea is that kernel functions enables us to work in the feature space without having to map the data into it.

Name of Kernel	$K(\mathbf{x}_i, \mathbf{x}_j) =$
Gaussian RBF	$\exp[-\frac{\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2}{c}]$
Polynomial	$((x_i.x_j) + c)^d$
Power Exponential (PE)	$\exp[-(\frac{\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2}{r^2})^\beta]$
Hyperbolic tangent or *Sigmoidal*	$\tanh[a(\mathbf{x}_i \cdot \mathbf{x}_j) + b]$
Cauchy	$\frac{1}{1+\frac{\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2}{c}}$
Inverse multi-quadric	$\frac{1}{\sqrt{\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2 + c^2}}$

There are many other kernel functions, such as spline functions that are used for support vector machines. In addition, many kernels have been developed for specific applications. However, in general it is very difficult to decide which kernel function is best suited for a particular application. This is an open problem among others.

5 Information Complexity

The choice of the best mapping function is not so simple and automatic. Presently a valid method for selecting the appropriate kernel function does not exist in the literature. Here, we propose to use the information complexity criterion of Bozdogan (1988, 1990, 1994, 2000, 2004) as our model selection index as well as our criterion for feature variable selection.

Since in the kernel methods make use of orthogonal and highly sparse matrices, in this paper we propose the modified entropic complexity of a covariance matrix.

Under a multivariate normal model, the maximal information-based complexity of a covariance matrix $\hat{\Sigma}$ is defined by

$$\begin{aligned} C_1(\hat{\Sigma}) &= \frac{s}{2}\ln\left(\frac{tr(\hat{\Sigma})}{s}\right) - \frac{1}{2}\ln\left|\hat{\Sigma}\right| \\ &= -\frac{1}{2}\ln\left(\prod_{j=1}^{s}\left(\frac{\lambda_j}{\overline{\lambda}_a}\right)\right) \\ &= \frac{s}{2}\ln\left(\frac{\overline{\lambda}_a}{\overline{\lambda}_g}\right) \end{aligned} \tag{18}$$

where $\overline{\lambda}_a = 1/s\sum_{j=1}^{s}\lambda_j \equiv tr(\hat{\Sigma})/s$ is the arithmetic mean of the eigenvalues (or singular values) of $\hat{\Sigma}$,and $\left|\hat{\Sigma}\right|^{1/s} \equiv \overline{\lambda}_g = \left(\prod_{j=1}^{s}\lambda_j\right)^{1/s}$ is the geometric mean of the eigenvalues of $\hat{\Sigma}$, and $s = rank(\hat{\Sigma})$.

- Note that $C_1(\hat{\Sigma}) = 0$ only when all $\lambda_j = \overline{\lambda}_a$.
- $C_1(\cdot)$ is scale invariant with $C_1(c\hat{\Sigma}) = C_1(\hat{\Sigma})$, $c > 0$. See, Bozdogan (1990).

Under the orthogonal transformation T, the maximal complexity in (18) can be written as

$$C_1^*(\hat{\Sigma}) = -\frac{1}{2}\sum_{j=1}^{s}\ln(s\lambda_j) \tag{19}$$
$$\cong \frac{1}{4}\sum_{j=1}^{q}(s\lambda_j - 1)(s\lambda_j - 3) - \frac{1}{6}\sum_{j=1}^{q} O\left(s\lambda_j - 1\right)^3,$$
$$0 < \lambda_j < \frac{2}{s},\ j = 1, 2, \ldots, s$$

where $O(\cdot)$ denotes the order of the argument and the Taylor series expansion of $\ln(s\lambda_j)$ used in (19) is about the neighborhood of the point

$$\lambda_1 = \lambda_2 = \cdots = \lambda_s = \frac{1}{s}. \tag{20}$$

At the point of eigenvalue equality $C_1^*(\cdot) = 0$ with $C_1^*(\cdot) > 0$ otherwise. See, Morgera (1985, p.610).

We note that (19) is only one possible measure of covariance complexity. Any convex function $\phi(\cdot)$, like $-\ln(\cdot)$, whose second derivative exists and is positive, may be used as a complexity measure, i.e.,

$$C_\phi^*(\cdot) = c\sum_{j=1}^{q}[\phi(\lambda_j) - \phi(\frac{1}{q})] \tag{21}$$

leads to an entire family of complexity measures, where c is a constant.

With this in mind, van Emden (1971, p. 63, eq. 311) suggested a second measure of complexity of a covariance matrix based on the Frobenius norm given by

$$C_F(\hat{\Sigma}) = \frac{1}{s}\left\|\hat{\Sigma}\right\|^2 - \left(\frac{tr\hat{\Sigma}}{s}\right)^2 \tag{22}$$

where $\left\|\hat{\Sigma}\right\|^2 = tr(\hat{\Sigma}'\hat{\Sigma})$, the square of the Frobenius norm of $\hat{\Sigma}$. In terms of the eigenvalues (or singular values), $C_F(\hat{\Sigma})$ reduces to

$$C_F(\hat{\Sigma}) = \frac{1}{s}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2. \tag{23}$$

Note that $C_F(\cdot) \geq 0$ with $C_F(\cdot) = 0$ only when all $\lambda_j = \overline{\lambda}$. Hence $C_F(\cdot)$ measures the absolute variation in the eigenvalues and it is translation invariant. That is, $C_F(\hat{\Sigma} + kI)) = C_F(\hat{\Sigma})$.

Since we can approximate $C_1(\hat{\Sigma})$ as

$$C_1(\hat{\Sigma}) \cong \frac{1}{4}\sum_{j=1}^{s}(\frac{\lambda_j - \overline{\lambda}_a}{\overline{\lambda}_a})^2, \tag{24}$$

in terms of the eigenvalues (or singular values) λ_j, $j = 1, 2, \ldots, s$, we can relate $C_1(\hat{\Sigma})$ to the Frobenius norm characterization of complexity $C_F(\hat{\Sigma})$ of $\hat{\Sigma}$ (Bozdogan, 1988) by introducing $C_{1F}(\hat{\Sigma})$ given by

$$
\begin{aligned}
C_{1F}(\hat{\Sigma}) &= \frac{s}{4}\frac{C_F(\hat{\Sigma})}{(\frac{tr(\hat{\Sigma})}{s})^2} = \frac{s}{4}\frac{\frac{1}{s}\left\|\hat{\Sigma}\right\|^2 - \left(\frac{tr\hat{\Sigma}}{s}\right)^2}{(\frac{tr(\hat{\mathcal{F}}^{-1})}{s})^2} \\
&= \frac{s}{4}\frac{\frac{1}{s}tr(\hat{\Sigma}'\hat{\Sigma}) - \left(\frac{tr\hat{\Sigma}}{s}\right)^2}{(\frac{tr(\hat{\Sigma})}{s})^2} \\
&= \frac{s}{4}\frac{1}{s\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2 \\
&= \frac{1}{4\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2.
\end{aligned}
\tag{25}
$$

We note that $C_{1F}(\cdot)$ is a second order equivalent measure of complexity to the original $C_1(\cdot)$ measure. Also, we note that $C_{1F}(\cdot)$ is scale-invariant and $C_{1F}(\cdot) \geq 0$ with $C_{1F}(\cdot) = 0$ only when all $\lambda_j = \overline{\lambda}$.Also, $C_{1F}(\cdot)$ measures the relative variation in the eigenvalues.

When it is assumed that the covariances are common between the classes or groups, we define ICOMP in KDA given by

$$
\begin{aligned}
ICOMP(\hat{\Sigma}_W) &= np\log 2\pi + n\log|\hat{\Sigma}_W| + np + 2C_{1F}(\hat{\Sigma}_W) \\
&= np\log 2\pi + n\log|\frac{1}{n}\hat{S}_W^{\Phi}| + np + 2\left[\frac{1}{4\overline{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \overline{\lambda}_a)^2\right]
\end{aligned}
\tag{26}
$$

In our numerical results, however, it suffices just to use and score $C_{1F}(\hat{\Sigma}_W)$ by itself to choose the optimal kernel function in KDA in the next section. In the literature cross-validation based criteria have been used. These type of criteria due to the high dimensionality of the feature space are too time-consuming. Our approach shortens the model selection time.

6 A Numerical Example

In this section we illustrate our results using the binary classifier trained by the KDA on Ripley's (1994) two dimensional toy data of $n_{trn} = 250$ training observations. Then, the classifier that is found is evaluated on the testing data of Ripley which has $n_{test} = 1000$ observations using support vector classifier (svmclass) to classify the input vector $\mathbf{x}$. We obtain both the training and

test error. Our results are based on the modified version of the STPRTool Matlab Modules of Franc and Hlaváč (2004). In our computation, we use *svds* function in Matlab to find the singular values of the large sparse within-group covariance matrix $\hat{\Sigma}_W$. We experimented with our results by retaining $k = 4$, and 15 largest singular values of $\hat{\Sigma}_W$ and scored the information-theoretic complexity $C_{1F}(\hat{\Sigma}_W)$ for each of the alternative kernel functions. The results from this experiment are summarized in Table 1 below. We report the results for $k = 4$ largest singular values only. The results up to the 15 largest singular values of $\hat{\Sigma}_W$ are the same in terms of the ordering of the complexity $C_{1F}(\hat{\Sigma}_W)$.

Kernel Function	Training Error	Test Error	$C_{1F}(\hat{\Sigma}_W)$
Linear	17.20%	14.20%	5.0276
RBF	13.60%	9.60%	4.5207
Polynomial [2 1]	15.60%	9.20%	4.6401
Sigmoid [2 1]	12.00%	9.00%	4.0747*

Table 1. Results KDA using different kernels functions and SVM Classifier.

Note that the regularization parameter μ was set to a small value 0.001, and the regularization constant C was set to 10. Looking at the above table we see that sigmoid kernel function seems to be a better choice based on the minimum value of the complexity measure $C_{1F}(\hat{\Sigma}_W)$ for this data set with better training and test error percentages. The visualization of the classifiers as SVM classifiers are shown in Figures 1, and 2.

7 Conclusion and Future Work

In this sort paper, we introduced the information-theoretic complexity $C_{1F}(\hat{\Sigma}_W)$ as a new method for model selection in choosing the optimal kernel function in kernel discriminant analysis (KDA). We showed our results on a toy benchmark data set of Ripley to evaluate the performance of the optimal classifier based on the choice of the kernel function. Our method shortens the model selection time over the more time-consuming cross-validation method.

The future work in this direction will involve several important problems in automating the choice of the regularization constant C, the regularization parameter μ and to study their effect on the classifier across different benchmark data sets and show the generalization ability of this new method. Our results, will be applied to real micro data mining data sets and the results will be reported elsewhere.

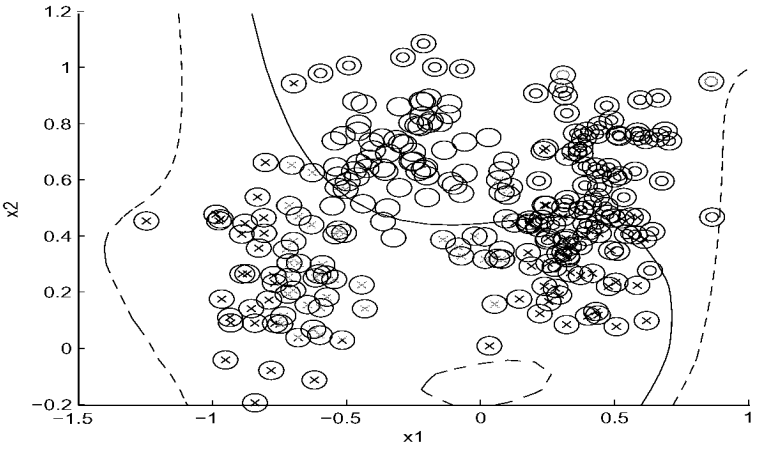

Fig. 1. Pattern of Ripley Training Data in the Feature Space.

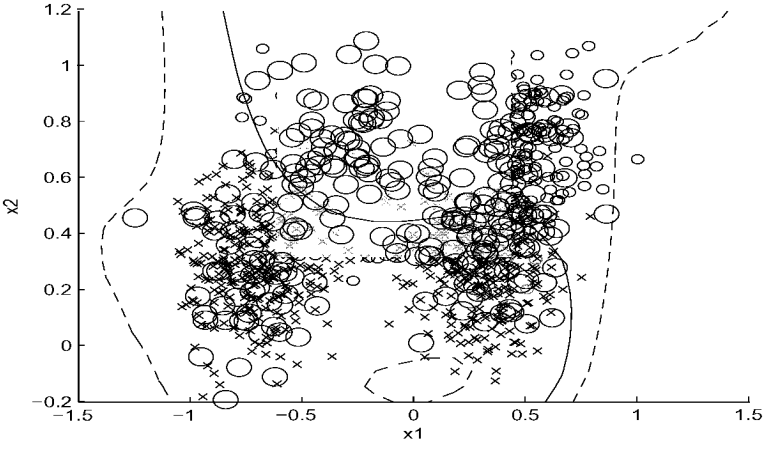

Fig. 2. Pattern of Ripley Test Data in the Feature Space.

Acknowledgements

The authors would like to express their thanks to Professor Isabella Morlini of the Università di Modena e Reggio Emilia, Italy for kindly formatting the paper, and to Dr. Russ Zaretzki of the Department of Statistics, Operations, and Management Science at the University of Tennessee, Knoxville, Tennessee, for reading and commenting on this paper.